DOCUMENT RESUME

ED 174 679 TM 009 651

AUTHOR Powills, Judith A.; And Others
TITLE Holistic Essay Scoring: An Application of the Model for the Evaluation of Writing Ability and the Measurement of Growth in Writing Ability Over Time.
PUB DATE Apr 79
NOTE 50p.; Paper presented at the Annual Meeting of the American Educational Research Association (63rd, San Francisco, California, April 8-12, 1979)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Age Differences; *Composition (Literary); Cost Effectiveness; *Essays; *Essay Tests; Evaluation Methods; Grades (Scholastic); Inservice Programs; Junior High Schools; Program Evaluation; *Rating Scales; Scores; *Scoring; Student Improvement; Test Reliability; Test Validity; Timed Tests; *Writing Skills
IDENTIFIERS *Holistic Evaluation; Interrater Reliability

ABSTRACT

Language arts teachers participated in an inservice program, The Writer's Clinic, and evaluated their students' improvement in writing ability using holistic essay scoring techniques. Seventh and eighth grade students were administered an essay composition pretest in December; the same essay topic was given as the post test in May. Students were allowed 30 minutes to complete the composition. Scores were assigned to the essays by 44 Writer's Clinic participants. Sample student papers were selected to illustrate the scoring standards, which included ratings of one through four. Readers assigned a score representing their impression of the total paper. Three teachers rated each student paper, without being aware of the scores assigned by other raters. Of 5,788 papers which were read and scored, 258 (4.46%) had discrepancies in scoring which were resolved by clinic staff. Interrater reliability ranged from .77 to .84, and readers were able to score an average of 64 papers per hour per rater. As hypothesized, differences were found between seventh and eighth graders, and between pretests and post tests. Validity of the scoring method was also supported by the relationship of essay scores to separate teacher ratings of students' writing abilities and to grades in composition courses. (GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED174679

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Holistic Essay Scoring: An Application of the Model for the Evaluation of Writing Ability and the Measurement of Growth in Writing Ability Over Time

by

Judith A. Powills
Educational Testing Service

Roberta Bowers
Indianapolis Public Schools

Gertrude Conlan
Educational Testing Service

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Judith A. Powills

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

April, 1979

Paper presented at the
Annual Meeting of the American Educational Research Association
San Francisco, CA: April 8-12, 1979
Session 3.23

TM009 651

THE PROBLEM

Providing teachers with information about the teaching of writing has received extensive interest. Such informational resources are applied within the classroom instructional framework; determination of the value and effectiveness of these informational resources as applied to a writing program requires product data available through the use of assessment instruments and methodologies. Such instruments and methodologies provide a key for decision making and the application of evaluation findings. One such assessment technique applied to the evaluation of writing ability is the holistic method of scoring student composition. If this method of analysis is to be of maximum effectiveness in terms of program evaluation, it is necessary to observe the method and its application in a model case study. To that end, this study investigated the appropriateness and utility of using the holistic essay scoring technique in order to evaluate a writing program and to measure growth in writing ability over time.

Examining a piece of student writing for its whole effect provides one means of analysis. Holistic scoring involves judging a composition for the total impression it creates, rather than judging individual components of writing including for example, spelling, punctuation, and organization. A single score is given for the complete paper rather than a series of scores on particular aspects of writing skill. Scorers do not ignore separate facets of writing; they however, consider each factor as it relates to the whole. Performance on one facet does not control the judgment of performance on the whole. Through the method, teachers view writing as an integrative process; the various components

of writing are considered simultaneously and teachers are encouraged to view teaching writing as more than the teaching of parts.

This study sought to determine the utility and outcomes of a holistic essay scoring in-service session for teachers of writing at the seventh- and eighth-grade levels. The compositions scored during the session included pretest and posttest essays. The problems examined for this study were: 1) whether or not the in-service scoring session complied with the theoretical model of holistic essay scoring: judging a paper in relation to other papers within the population rather than against a preconceived ideal; 2) whether or not one could have confidence in the obtained scores: maintaining standards set by the readers and achieving consistency in scoring the essay papers as evidenced by reasonably high interrater reliabilities; 3) whether or not writing ability improved over progressive grade levels: finding significant differences in test performance of seventh- and eighth-graders; 4) whether or not writing improved over a short period of time: finding significant differences between pretest scores (test written in December) and posttest scores (test written in the following April); 5) whether or not the writing program implemented by the study schools was effective: students of program participating teachers performing better at posttest time than students of non-program teachers; and finally, 6) whether or not the obtained essay scores were valid measurements of writing ability.

Need for the Study

The application of the holistic method of scoring essays is a response to the demand for the direct measurement of writing ability.

Writing abilities have most often been assessed by indirect measures such as standardized tests in a multiple-choice format. The obtained high reliability and validity coefficients, as well as high correlations with direct measures of writing, lend credence to the use of such indirect measures. Nevertheless, teachers of writing have proclaimed that indirect measures provide analogous assessments of the students' ability to write, at best. If learning to write compositions is the ability to be measured, then actual composition, it is said, should be assessed rather than some analogous behavior.

Direct measurement of writing surely, will continue to be sought particularly within the classroom; it will not give way totally to multiple-choice tests of writing skill. In fact, given a highly reliable and valid means for scoring composition, this direct measurement of writing may have stronger implications within the schools and within the district, for large-scale testing programs, and for the analysis of writing abilities for placement, admission, and employment purposes.

The need for this study is inherent in providing a quick, efficient, reliable and valid means to evaluate writing. Use of the holistic method by teachers of writing from varied backgrounds and different schools, their reactions to the technique, the results and implications of the scores obtained, can provide a basis for educators interested in the direct measurement of writing. Since the holistic method of scoring composition may provide a means of helping teachers to improve their teaching of writing, and of helping administrators to assess the writing skills of students in their schools, it would be useful to study the use of the method by teachers of writing in order to evaluate composition, to measure growth in writing over time, and to evaluate a program designed to assist students in refining their knowledges and skills in

the writing process. The methodologies and findings of this study may prove useful to other teachers and administrators with programs, goals, and objectives designed to improve student writing abilities.

The Study's Relationship to the Indianapolis Public Schools' Writer's Clinic

The Writer's Clinic was an in-service program for junior high school teachers implemented by the Indianapolis Public Schools and sponsored by a grant from the Lilly Endowment. Designed to aid teachers in enhancing the writing skills of their seventh- and eighth-grade students, the program was offered to all junior high teachers on a voluntary basis. Forty-eight Language Arts teachers of grades seven and eight, participated and committed themselves to complete a series of activities during the course of thirteen in-service days of the Writer's Clinic in a one academic year period. The teacher contracts included: writing once a week with the students; guiding students in self-editing and self-proofreading; maintaining student folders containing writing; directing students in completing check sheets for the first composition of the year and the third composition of each six weeks period; submitting one flawless piece of prose from each student at the end of the project; and communicating with parents regarding the activities of the clinic and individual student progression. Ultimately, in the final two days inservice for the 1977-78 Writer's Clinic, pretest and posttest essays would be evaluated holistically not only to provide teachers with experience in the method, but also to evaluate the program's effectiveness in terms of student growth in writing.

The Writer's Clinic for teachers began with an intensive series

of five work sessions during the summer. Session leaders included Indianapolis Public School professional and administrative staff (Ms. Roberta Bowers, Writer's Clinic Clinician and Ms. Helen Cartwright, Supervisor of Language Arts) as well as invited resource leaders from other institutions and organizations (Mr. Harvey Jacobs, Editor, The Indianapolis Star; Dr. H. Thompson Filmer, Professor of Education, University of Florida, Gainsville). Areas of concern for the initial in-service activities included: the starting point and setting of the project; composition through literature; our language today; uses of expository paragraphs in speech and debate and in reading units; the development of unit themes, from unit themes to thesis statements; thought and language models in organizing experiences; and evaluation by writing, editing content, and proofreading. The primary focus of these early in-service days was on using a variety of models in the teaching of composition.

Two in-service work sessions including the topics of a writing experience from planning to final copy, editing and evaluating, sharing successes, and the basics of writing, were held in October and November. The second clinic was conducted by Dr. Hans Guth, coauthor of the series of English textbooks, American English Today. Dr. Guth demonstrated how composition could be taught through a progressive sequence from a word to a sentence to a paragraph and finally extending to a complete theme. The third clinic was led by Ms. Roberta Bowers. In this session, participants were led through a lesson in descriptive writing. The outcome of this lession was participant-written essays that were compiled into booklet form.

In January, the Clinic included two work sessions directed by

Dr. Michael Flanigan, Director of First Year Studies, Indiana University, Department of English. Topics included: a nine-step writing process; the skills of writers and the skills of teachers; and students as their own editors. Using specified models, particularly a process approach to writing which concentrated on the skills which writers possess and what writers actually do, participants focused on structure and audience and analyzed their own writing behaviors.

One work session comprised the fifth Writer's Clinic. Staff from the Indianapolis Public Schools including an evaluation consultant, Mr. Carl Hines, led the session. Participants discussed writing as a form of self expression and covered the topics of reading, reflecting, and writing; preparing an entry for a Writer's Fair; and the evaluation of writing. The last segment of this work session in March, involved a sharing of successful experiences among the participating teachers.

The sixth Writer's Clinic meeting was conducted by Dr. Harrison J. Means, Associate Professor of English, Curriculum and Foundation Division, Cleveland State University. Dr. Means suggested strategies for teaching writing to slower students and discussed miscue analysis of writing and what it tells the writer, the diagnostic/prescriptive model format for teaching composition; criterion-referenced evaluation strategies, and sentence combining strategies.

The seventh and final meeting of the Writer's Clinic was held in May for two days. Teachers participating in the clinic activities, having incorporated their clinic experiences into their own classroom instruction, along with program administrators, were in need of an efficient method for evaluating the program in terms of its effectiveness as demonstrated by growth in the students' writing ability. Pretest

and posttest essay papers had already been submitted for evaluation purposes. This concern was the focus of the last clinic where the Educational Testing Service conducted a two-day holistic essay scoring session. Clinic participants were trained in the method of holistic scoring and proceeded to read and score approximately 6,000 student compositions from two groups: experimental (students whose teachers had participated in the Writer's Clinic program) and comparison (students whose teachers had not participated in the program, but who were identified to represent characteristics of the experimental group). The methodology of this scoring session and the scores assigned to the students' compositions by the readers were the prime focus of this study. Evaluation results of this study and other assessment data were to provide a measure of the Writer's Clinic program effectiveness for the Indianapolis Public Schools.

Limitations of the Study

Because this study was subject to several limitations, caution in interpreting the results of the investigation is necessitated.

The study sample was not randomly selected. Experimental students who submitted pre- and posttest essays were seventh- and eighth-grade students of teachers who had participated in the Writer's Clinic. The Writer's Clinic, in its first year of operation, was volunteer-participation-based. Therefore, the lack of sample randomness was a necessary limitation; it would neither have been feasible or desirable to have teachers assigned at random to participate in the clinic. It is also important to point out that the teachers had a written contract to comply with the goals, objectives, and activities outlined for the

program and described earlier. This commitment, as well as participating in thirteen actual days of inservice, emphasized that the volunteering teachers were highly motivated. They can not be necessarily assumed to be a typical group of teachers.

Comparison group students were to be selected so as to be as similar as possible to student participants in the Writer's Clinic with respect to variables such as age, grade, sex, class, and school; while a one-to-one match was not necessary, the overall group profiles should have corresponded. The comparison group was formulated to match characteristics of seven hundred randomly selected experimental students, however, the representativeness of the comparison group was less than ideal. This was due partially to the fact that comparison group students were selected from intact classes. Again, without working with intact classrooms, selection of any comparison students would not have been feasible. Additionally, comparison students were selected after the pretest essay had been administered to experimental students. While the comparison group students did respond to the same posttest essay assignment at the same time as experimental students, pretest essays were not received from comparison students. Keeping the design limitations in mind, analyses including comparisons between treatment and non-treatment group posttest scores were conducted.

METHODOLOGY

Questions to be Answered

The questions to be investigated included:

1. Does the case study fit the theoretical model of holistic essay scoring?

2. Is consistency in scoring the essay papers achieved as evidenced by determination of interrater reliabilities?

3. Is there a significant difference in test performance of seventh- and eighth-graders as indicated by the obtained total test scores?

4. Is there significant growth in writing ability over time, from pretest to posttest?

5. Is the Writer's Clinic inservice program for teachers effective as indicated by a significant difference in post-test scores obtained by the comparison and experimental groups?

6. Are the pretest and posttest obtained scores valid measurements of writing ability?

Differences Between Treatments

Writer's Clinic participants were interested in directly evaluating student writing ability. The program clinician developed an essay topic to be administered to all seventh- and eighth-grade students whose teachers volunteered for clinic participation. This essay assignment served as the pretest for the evaluation of writing abilities. The students were to respond to the same essay topic at the end of the school year. This second assignment constituted the posttest for the evaluation. A pretest-posttest design for the evaluation of writing skill was therefore, implemented.

In order to determine the effectiveness the teacher inservice program had in terms of student writing ability, the study sample was comprised of two primary groups. The group taking both the pretest and posttest was the experimental group. Treatment for this group consisted of the classroom experiences incorporated by the students' teachers who had volunteered to participate in the Writer's Clinic program. The

'no treatment' group, comprised of students whose teachers did not participate in the Writer's Clinic formed the comparison group. Comparison group students, unlike experimental students, did not write in response to the pretest essay assignment. Both groups, experimental and comparison, were compared using t-tests of differences between posttest means.

Sample

The final sample for analyses included compositions written by 4,071 seventh- and eighth-grade students in the Indianapolis Public Schools. Eighty-eight compositions were considered invalid and were deleted from the analyses. Invalid compositions included those that were off the topic or were incomprehensible and those for which student identifying information was missing. Of the 4,071 study students, 648 formed the comparison, posttest only subgroup. Comparison group students came from eight different schools. Of the 3,623 experimental students, 1,717 had taken the pretest and posttest (matched pairs); 804 had taken the pretest only; and 902 had taken only the posttest. The pretest only and posttest only experimental groups resulted due to mobility or absence on the day of test administration. Experimental students came from 31 different schools. Fifty teachers at the seventh- and/or eighth-grade teaching levels had volunteered to participate in the Writer's Clinic program.

Data Collected

Data collected for the study included pretest and/or posttest scores assigned the student essays written for the Writer's Clinic evaluation. The scores were assigned during a holistic essay scoring session

conducted in May of 1978. This session represented the final Writer's Clinic activity for the school year. Each paper was read and scored by three different readers during the session. A total of 44 Writer's Clinic participants assisted in the scoring. After scoring discrepancies were resolved, the three individual scores given each paper were summed to derive a total paper score. Student identifying information including school and grade level, were collected and matched with first, second, third, and total reading scores.

In order to obtain some measure of concurrent validity, the essay scores were correlated with a criterion derived from class composition grades and teacher ratings of overall student writing ability. Each of the Writer's Clinic teachers who participated in the scoring session were asked to provide student information to reflect the criteria for students in one of their classes. A random assignment schedule was arranged to determine the classes for which criteria information would be provided. The teachers were given a form and were asked 1) to enter the names of all students in the identified class, 2) to enter the grades given the essay assignments for each month from December to May corresponding to each student, and 3) to rate each student on his/her overall writing ability in December and again in May using a scale of one (low) to four (high). Forms were received from only 15 of the 44 teachers. Nine of the 15 teachers taught students at the seventh-grade level and six taught students at the eighth-grade level. Class essay grade data were collected for 396 students across the classes and grade levels; 244 seventh-graders and 152 eighth-graders. These data were matched with the pretest and/or posttest scores obtained during the essay scoring session.

Three of the 15 teachers who completed the form did not enter the overall ratings of writing ability for their students but did supply essay grades; two eighth-grade classes and one seventh-grade class lacked the overall rating data. Overall ratings were obtained for a total of 300 students.

In order to determine the relationships between composition grades and essay test scores and teacher ratings and essay test scores, only those students having both pretest and posttest essay scores were included in the analysis (N=215). Of the 215 students, 135 were eighth-graders and 80 were seventh-graders. The teachers reported composition grades but did not rate 57 of these students in terms of overall writing ability in December and May (three classes); overall ratings were received for 158 of the 215 students.

The number of composition grades submitted varied by class and by student. No student had more than five composition grades reported within any one month (December through May). Because the grades were separated into two periods relative to the pretest period and posttest period (December to February and March to May) students having fewer than four reported grades within either of the two periods were deleted from analysis. Means for reported grades were calculated for both periods (pre- and post-) and it was determined that students with fewer than four reported grades within a three month period could bias the period mean.

The Holistic Essay Scoring Procedure

Activities Prior to the Scoring Session

Students were requested to respond to a direct measure of writing ability; they were to respond to a given topic by developing a composition theme. Concerned with measuring progress in writing effective-

ly, it was decided that two controlled writing experiences would be conducted, one in December (pretest) and one at the end of the school year (posttest in May). The controlled writing assignment incorporated several procedures: there were no prewriting activities; the assignment was given succinctly and a certain amount of time was allowed for the writing (30 minutes); students were instructed to use standard junior high school format; students were allowed to revise and rewrite during the allotted time; teachers were not able to assist except to clarify the directions; the assignments were not to be corrected or graded by the teachers; and teachers were given a limited time period in which to have students complete the assignment.

The difficulty level of the topic was controlled for by using the same topic for the two test administrations. Research has shown that different essay topics do in fact, represent different difficulty levels (Godshalk, 1966). With the variability of different topics controlled for, reading standards would remain consistent between pretest and posttest compositions. The practice effect was considered to be minimal because of the four month time lapse between the test administrations.

The essay topic selected for the pretest and posttest essay assignments was developed by a professional writing clinician of the Indianapolis Public Schools. Established guidelines for writing appropriate essay topics were followed. Student pretest and posttest compositions were submitted to the Writer's Clinic clinician for processing previous to the scoring session. In preparation for the reading, all papers were randomly mixed among the classrooms, teachers, grade levels, schools and test type (pretest or posttest).

Anonymity of student responses was assured by secluding all identifying information (name, grade, class, teacher, school, test date). Removeable labels covered this identifying information. By assuring writer anonymity, reader bias was controlled for; the reader would be unaware of who the writer was, what class, grade or school the writer was from, and whether or not the paper was written for the pretest or posttest administration.

Prior to the scoring session, workshop leaders read the majority of pretest and posttest student compositions in order to select sample papers that would be used for training in setting scoring standards. Sample papers were selected to represent the various levels (ranges) of population writing ability from excellent to average to poor. The score scale that was to be used for the session ranged from one to four (low to high, respectively). Sample papers included those judged to be representative of each score. The number of samples were normally distributed across scores; there were more samples scored by the workshop leaders as 'two' and 'three' than samples scores as 'four' and 'one.' Other sample papers selected were those that might present particular problems to the readers or those that might present some other interest to the readers.

Unique score code cards were developed for the readers to assure independent judgment during the reading. These score codes were established to keep the judgments of each of the three readers of each paper independent. Each reader was to receive a unique score code so that when the second reader read a paper, he or she would be unable to interpret the score assigned by the first reader, and likewise for the third reading.

Papers were inserted into manila file folders, approximately 20 per folder. Readers would score a complete folder before scoring papers in another folder.

Scoring Session Procedures

Introduction to holistic scoring. Forty four teachers of the Indianapolis Public Schools were in attendance at the scoring session. The first task for the workshop leaders was to introduce the method to the readers. Readers were told that holistic scoring is impressionistic scoring. The paper is read and scored for the total impression it creates rather than for particular aspects of writing skill such as punctuation, organization, diction, and/or spelling. The score scale of 'one' to 'four' was explained to the readers. In a scale which has only four points, there is obviously no middle score. One of the reasons why the score scale works as well as it does, is that readers are forced away from a middle or uncommitted score. Readers were told they would make two judgments for each paper: first, a decision was required as to whether the paper was in the upper half or lower half (above average or below average); then, readers were to decide if the paper was good enough to be rated as the highest score, or poor enough to be rated as the lowest score. Papers were to be judged in relation to the other compositions in the population rather than against a preconceived ideal. How the papers would be judged would be determined through standards set by the group. Readers were to take a positive approach to rating the compositions. They were asked to concentrate on what the student had accomplished rather than on what the student had failed to do or had done badly. Readers were also reminded that the students had only 30 minutes in which to respond to the topic, and that they were unable

to use aides such as a dictionary.

Setting the standards. Following the introduction to holistic scoring, scoring standards were set through the examination of the selected sample papers. Sample papers were distributed to readers one-by-one. Participants were asked to read the composition quickly, and to score it on a scale of 'one' to 'four' for the total impression it created. The workshop leader then asked for an open vote on how many people had assigned a score of 'four,' of 'three,' of 'two,' or of 'one,' to the paper. Group votes were counted, tallied, and recorded on a blackboard visible to the entire group. A total of six sample papers representing the range of writing abilities were distributed, scored, and openly voted upon in this manner. After each of the six samples had been scored, readers were asked to reexamine these papers and to put them in rank order as they related to each other. This was to reinforce the stipulation that each paper be read and judged against the other papers rather than against an ideal. Participants were then free to discuss the characteristics of any particular paper; readers who had assigned scores of 'four,' 'three,' 'two,' or 'one,' discussed the reasons for assigning the respective scores.

Examination of sample papers one-by-one and discussion of them by the group continued until the readers had reached consensus on the standards for judging the papers. Discrepancies in scoring were discussed. A discrepancy occurred if all readers had not judged an individual paper alike. For example, if the majority of readers placed a paper in the above average (three to four) category, and other readers placed the paper in the below average (one to two) category, discussion was called for. Discussion of the paper assisted the low scorers to view the paper as the majority viewed the paper. Divergence from the

standards on the part of any reader was corrected. Readers were to conform to the standards set by the group. Whether or not consensus was reached was determined by the examination of the count of scores as recorded on the blackboard. After a series of papers had been read, consensus was achieved; the majority of readers assigned the same score to each sample paper. It is important to note that the workshop leader did not impose standards on the group; rather, the group set its own standards.

The actual scoring. After standards for the reading were set with consensus in scoring achieved, each reader was given his or her own unique scoring code. Readers were instructed as to the use of this score code, and how to write the code on the paper. Each reader received a folder of approximately 20 papers and was requested to read each paper in the folder and to score it according to the standards set during the training. Upon completion of a folder, the reader was given another. During the first round of reading, each folder was checked by the workshop leaders to assure that scoring standards had been maintained, and that the readers were adhering to these standards in their scoring. All papers were read once before distribution for the second round of reading.

Before papers were distributed for the second reading, they were taken out of the original folders and randomly mixed. New folders were compiled which included papers read and scored by many different readers during the first reading. This mixing system was incorporated so that the second reader would receive papers already read and scored for the first time by several readers. The same mixing procedure was implemented after all papers had been read twice, before distribution for the third reading.

At the end of the first day, all papers had been read at least once, and a large proportion had been read two times. A mini-training session took place at the beginning of the second day to assure that the standards set during the first day of the scoring session were maintained on the second day. By the middle of the afternoon on the second day of scoring, all papers had been read and scored three times by three different readers. Papers were redistributed to the readers to convert the score codes into actual scores. Readers also totalled the three individual scores to obtain the total score.

After the Scoring Session

Following the reading, papers were examined by the clinician of the Writer's Clinic to resolve discrepancies in scoring. Discrepancies were defined as non-adjacent assigned scores. The clinician resolved each discrepancy so that the three scores were in fact at least adjacent. A total of 258 papers of the 5,788 papers read and scored had discrepancies (4.46 percent). The resolution of discrepancies involved changing one, two, or three of the assigned scores. Most papers had only one score changed (84.11 percent); fewer needed a change for two of the three scores (15.12 percent); and only two papers (.78 percent) required a change for all three scores. Most changes (35.65 percent) were made to the first reading score. Only three cases of the 258 discrepancies resulted in no change to the total score. Total scores were decreased for 123 (47.67 percent) of the total discrepancies and were increased for 132 (51.16 percent) of the total. The range of changes to the total score was from minus three score points to plus five score points.

The labels covering the identifying student information were removed so that each paper could be coded according to the following variables: student identification number; school code; grade; pretest first reading score, second reading score, third reading score, and total pretest score; posttest first reading score, second reading score, third reading score, and total posttest score. These variables were coded, keypunched and verified. All variable data were then computer edited for valid values. After errors found by the computer edit were corrected, a five percent random sample of cases (N=200) was generated to determine coding errors. The analysis identified a .4 percent item coding error rate.

RESULTS

The Theoretical Model of Holistic Essay Scoring

Theoretically, the scores for a large-scale holistic essay reading should approximate the normal curve. This is due primarily to the fact that the essays for this type of scoring are judged in relation to other papers in the population; they are not judged against a preconceived ideal. Therefore, one would expect that within one reading the number of papers given scores of 'two' and 'three' would be higher than those given scores of 'one' and 'four.' One would also expect that the number of papers given scores of 'one' and 'four' would be similar as would the number of papers given scores of 'two' and 'three.'

Statistics calculated to determine the degree to which a distribution of cases approximates a normal curve include the third and fourth moments; skewness and kurtosis, respectively. Skewness will equal zero when the distribution represents a bell shaped curve which is completely symmetrical. Kurtosis will also equal zero when the distribution represents a perfect normal curve.

Table 1 shows the distribution of cases for each reading of all papers across all groups including obtained measures of skewness and kurtosis. The positive values of skewness indicate the deviations from perfect symmetry are such that there is a clustering to the left of the mean with extreme values to the right. Negative kurtosis values, across all readings, show the distributions to be flatter than a normal distribution.

Table 1

Distribution of Raw Scores by Reading
(N=5,788)

Score	First Reading		Second Reading		Third Reading	
	N	%	N	%	N	%
1	854	14.8	664	11.5	615	10.6
2	2,381	41.1	2,541	43.9	2,733	47.2
3	1,781	30.8	1,963	33.9	1,919	33.2
4	772	13.3	620	10.7	521	9.0
Mean	2.43		2.44		2.41	
Standard Deviation	.90		.83		.80	
Skewness	.16		.16		.22	
Kurtosis	-.73		-.53		-.38	

In terms of skewness, the first and second readings appeared to best approximate the normal curve. The third reading, on the other hand, appeared to best approximate the normal curve in regard to kurtosis. While specific tests of significance were not performed, the statistics indicate that the null hypothesis that the distribution of cases does not approximate the normal curve can be rejected with some certainty.

Values of skewness and kurtosis were also calculated for the distribution of total scores summed across the three readings. The range of total scores was three to twelve. This distribution as presented in Table 2, also indicates scores to cluster to the left of the mean with extreme values to the right, and that the distribution is flatter than that of a perfectly normal distribution; skewness of .22 and kurtosis of -.49.

Table 2

Distribution of Total Raw Scores
(N=5,788)

Score	N	Percent
3	208	3.6
4	334	5.8
5	640	11.1
6	1,145	19.8
7	973	16.8
8	793	13.7
9	777	13.4
10	414	7.2
11	282	4.9
12	222	3.8

Mean	7.27
Standard Deviation	2.17
Kurtosis	-.49
Skewness	.22

Interrater Reliability

The value of scores is dependent upon their objectivity and reliability. To the degree that qualified observers would assign different scores to the same paper, the measurement lacks objectivity and

utility. Before proceeding to analyses dependent upon assigned scores, it was necessary to establish the interrater reliabilities. Each student's essay was read three times by three distinct readers. Table 3 presents Cronbach's alpha coefficients of reliability for the reading of all student essays by study subgroups. The reliability coefficients ranged from .77 to .84. Clearly, the reading reliability was not so strong so as to suggest perfect agreement between readers. However, the obtained coefficients seemed substantial enough to proceed with the other study analyses.

Table 3

Alpha Coefficients of Reliability by Group and Grade

Group	Alpha
All experimental group pretests	
7th Grade	.80
8th Grade	.83
All experimental group posttests	
7th Grade	.82
8th Grade	.84
All comparison group posttests	
7th Grade	.77
8th Grade	.78
All posttests (comparison and experimental groups)	
7th Grade	.81
8th Grade	.83

Differences Between Seventh- and Eighth-Graders

To determine if differences existed between seventh- and eighth-grade students on pretest and posttest, various t-tests comparing sample total score means were performed by population subgroups. The t-tests included those for independent samples where cases were classified into two groups and a test of mean differences was performed. Such tests of significance were performed for the two distrinct groups; pretest and posttest (experimental and comparison groups).

The experimental group was comprised of two subgroups for the pretest: students having taken only the pretest (pretest only subgroup); and students having taken both the pretest and posttest (pretest/posttest matched pairs). The two subgroups could not be combined by pretest unless statistically, there was no difference between the pretest only seventh-graders and the pretest/posttest matched pairs seventh-graders, and likewise between the pretest only and pretest/posttest matched pairs eighth-graders. If differences within grade by subgroup were shown to exist, then comparisons between the grade levels could be analyzed only individually by subgroup. For all t-tests between independent groups, an approximation to t was computed based on separate variances where sample variances were unequal. t, based on the pooled variances was computed where sample variances were not unequal as indicated by the F-test of sample variances.

Tables 4 and 5 present one-tailed t-test results to determine differences on the pretest between the seventh- and eighth-grades. The stated null hypothesis was that the seventh-grade mean was equal to or greater than the eighth-grade mean.

Table 4

t-test Between Seventh and Eighth Grade Pretest Means
For the Experimental Pretest/Posttest Matched Pairs Subgroup
(N=1717)

Group	N	Mean	t	P
7th Grade	894	6.84	*4.81	.000
8th Grade	823	7.34		

*Approximation to t based on separate variances.

Table 5

t-test Between Seventh and Eighth Grade Pretest Means for
the Experimental Pretest Only Subgroup
(N=804)

Group	N	Mean	t	P
7th Grade	437	6.58	*5.28	.000
8th Grade	367	7.37		

*Approximation to t based on separate variances.

The null hypothesis that the seventh-grade mean is equal to or greater than the eighth grade mean for each of the two experimental subgroups having taken the pretest was rejected at the .000 significance level.

To determine whether differences existed between the seventh- and eighth-grade total score means for both pretest experimental groups combined, a two-tailed t-test was first performed between the seventh-graders of the pretest only subgroup and the pretest/posttest matched

pairs subgroup. The same test was performed between the means of both groups for the eighth-graders. Results are shown in Tables 6 and 7.

Table 6

t-test of Seventh Grade Pretest Means Between the Pretest Only and Pretest/Posttest Matched Pairs Experimental Subgroups (N=1331)

Group	N	Mean	t	P
Pretest Only	437	6.58	*2.26	.024
Pretest/Posttest Matched Pairs	894	6.84		

*t based on pooled variances.

Table 7

t-test of Eighth Grade Pretest Means Between the Pretest Only and Pretest/Posttest Matched Pairs Experimental Subgroups (N=1190)

Group	N	Mean	t	P
Pretest Only	367	7.37	* .27	.788
Pretest/Posttest Matched Pairs	823	7.34		

* t based on pooled variances.

The null hypothesis that there was a difference between the seventh-grade means of the pretest only and pretest/posttest matched pairs experimental subgroups was not rejected. The null hypothesis that there was a difference between the eighth-grade means of the pretest only and the pretest/posttest matched pairs experimental subgroups was rejected.

Since the seventh-grade means across both pretest experimental subgroups were statistically different, combining the seventh-graders from the pretest only and pretest/posttest matched pairs subgroups, and likewise for the eighth-graders of the experimental group would lead to misleading results. In summary, however, the eighth-grade mean was higher than the seventh grade mean for both pretest experimental groups as would be expected.

Tables 8 to 10 present one-tailed t-test results to determine differences on the posttest between the seventh- and eighth-graders; the stated null hypothesis that the seventh-grade mean is equal to or greater than the eighth grade mean for each posttest subgroup.

Table 8

t-test Between Seventh and Eighth Grade Posttest Means for the Comparison Group (N=648)

Group	N	Mean	t	P
7th Grade	179	6.78	* .54	.295
8th Grade	469	6.87		

* t based on pooled variances.

For the comparison group, the null hypothesis that the seventh-grade posttest mean is equal to or greater than the eighth-grade posttest mean was not rejected.

Table 9

t-test Between Seventh and Eighth Grade Posttest Means for the Experimental Pretest/Posttest Matched Pairs Subgroup (N=1717)

Group	N	Mean	t	P
7th Grade	894	7.35	*5.83	.000
8th Grade	823	7.96		

* t based on pooled variances.

For the experimental pretest/posttest matched pairs subgroup, the null hypothesis that the seventh-grade posttest mean is equal to or greater than the eighth-grade posttest mean was rejected at the .000 significance level.

Table 10

t-test Between Seventh- and Eighth-Grade Posttest Means for the Experimental Posttest Only Subgroup (N=902)

Group	N	Mean	t	P
7th Grade	505	7.48	* .78	.437
8th Grade	397	7.60		

*t based on pooled variances.

For the experimental posttest only subgroup, the null hypothesis that the seventh-grade posttest mean is equal to or greater than the eighth-grade mean was not rejected.

Two tailed t-tests were performed between the posttest seventh-grade means of the posttest only subgroup and the pretest/posttest

matched pairs subgroup, and also for the posttest eighth-grade means of both experimental posttest groups. This was done in order to determine if all experimental seventh-grade scores could be compared to all experimental eighth-grade scores. Results of the two t-tests are shown in tables 11 and 12.

While there was no difference for the seventh-grade posttest means between the two experimental posttest subgroups, a significant difference for the eighth-grade posttest means of the two subgroups was noticed. Therefore, combining across subgroups to test the difference between all experimental seventh-graders and all eighth-graders on the posttest would provide misleading results.

Table 11

t-test of Seventh-Grade Posttest Means Between the Posttest Only and Pretest/Posttest Matched Pairs Experimental Subgroups (N=1,399)

Group	N	Mean	t	P
Posttest Only	505	7.48	*1.06	.290
Pretest/Posttest Matched Pairs	894	7.35		

*t based on pooled variances.

Table 12

t-test of Eighth-Grade Posttest Means Between the Posttest Only and Pretest/Posttest Matched Pairs Experimental Subgroups (N=1,291)

Group	N	Mean	t	P
Posttest Only	397	7.39	*5.15	.000
Pretest/Posttest Matched Pairs	894	7.84		

*t based on pooled variances.

Growth in Writing Ability Over Time

Analyses to determine whether or not there was growth in writing ability over time (a significant difference between pretest and posttest scores, with posttest scores being higher) included only the pretest/posttest matched pairs experimental subgroup. This was due partially because in fact, there were differences in the pretest total scores between the pretest only and the pretest/posttest matched pairs subgroups, as well as differences in the posttest total scores between the posttest only and the pretest/posttest matched pairs subgroups. Additionally, the use of matched pairs for gain analyses receives attention in the literature.

The minimum total score for both pretest and posttest was three; the maximum total score for each essay administration was twelve. The maximum possible loss from pretest to posttest was nine points; the maximum possible gain from pretest to posttest was also nine points. Analyses to determine gain, loss, or no change from pretest to posttest were performed for each of grades seven and eight, as well as for both grades combined.

Table 13 shows the number of students and corresponding percentages of students who performed better on the posttest than on the pretest, who exhibited no change in total score from pretest to posttest, and who performed better on the pretest than the posttest for the seventh-grade, eighth-grade, and combined groups.

More students in both grades seven and eight performed better on the posttest than on the pretest than students either exhibiting no change or a negative gain. Students gaining comprised almost fifty percent of the group.

Table 14 presents the pretest to posttest change values for both grades individually and combined. The mean change value for seventh-graders was +.50 with a standard deviation of 2.0. The mean change value for eighth-graders was .21 with a standard deviation of .87. The mean change value for the total group was .56 with a standard deviation of 2.1.

Table 13

Distribution of Pretest to Posttest Gain Scores
(N=1,717)

		Positive Gain		No Gain		Negative Gain	
Group	Total N	N	%	N	%	N	%
7th Grade	894	444	49.66	174	19.46	276	30.87
8th Grade	823	413	50.18	170	20.66	240	29.16
Total	1,717	857	49.91	344	20.03	517	30.00

Table 14

Distribution of Pretest to Posttest Change Values

Change Value	Group					
	7th Grade		8th Grade		Total	
	N	%	N	%	N	%
+9	--	--	1	.1	1	.1
+8	--	--	--	--	—	—
+7	1	.1	3	.4	4	.2
+6	8	.9	11	1.3	19	1.1
+5	11	1.2	19	2.3	30	1.7
+4	44	4.9	41	5.0	85	5.0
+3	72	8.1	72	8.7	144	8.4
+2	137	15.3	126	15.3	263	15.3
+1	171	19.1	140	17.0	311	18.1
0	174	19.5	170	20.6	344	20.0
-1	131	14.7	116	14.2	247	14.4
-2	85	9.5	69	8.4	154	9.0
-3	46	5.2	37	4.5	83	4.8
-4	10	1.1	9	1.1	19	1.1
-5	2	.2	7	.8	9	.5
-6	1	.1	1	.1	2	.1
-7	—	--	--	--	--	--
-8	--	--	1	.1	1	.1

It must be noted that due to the regression effect, or regression towards the mean, students with the lowest pretest scores can appear to gain more than students with higher initial scores. Raw change or gain scores formed by subtracting pretest scores from posttest scores, as shown in Table 14 can lead to fallacious conclusions, because these scores are systematically related to random error of measurement. The mean gain scores as reported above do not seem substantial. A different picture of growth in writing ability over time, however, is apparent

when significance tests are performed to determine if the posttest means are reliably different from and greater than the pretest means. The difference in sample means for pretest and posttest is a better estimate of the mean difference than is the analysis of raw gain or change scores.

Correlated t-tests for paired samples were performed for seventh-grade, eighth-grade, and combined grades between pretest and posttest means. Results are shown in Tables 15 to 17.

Table 15

Correlated t-test Between Pretest and Posttest Means for Seventh-Graders (N=894)

Group	Mean	t	P
Pretest	6.84	7.51	.000
Posttest	7.35		

The null hypothesis that there is no difference between pretest and posttest means was rejected at the .000 significance level for seventh- and eighth-grades individually and combined.

Table 16

Correlated t-Test Between Pretest and Posttest Means for Eighth-Graders (N=823)

Group	Mean	t	P
Pretest	7.34	8.22	.000
Posttest	7.96		

Table 17

Correlated t-Test Between Pretest and Posttest Means
For All Matched Pair Students
(N=1,717)

Group	Mean	t	P
Pretest	7.08	11.13	.000
Posttest	7.64		

The Effectiveness of the Writer's Clinic Program

Comparison Versus Experimental Groups

Before posttest differences between the experimental (treatment) and comparison (no treatment) groups were analyzed, it was necessary to determine whether or not the two groups were essentially equivalent at pretest time. The comparison group was selected after experimental students had taken the pretest. Therefore, the comparison group students were administered the posttest only. Without the pretest for comparison students, a t-test between experimental and comparison means was not possible.

In order to compare the status of program participants following treatment to that of similar non-participants, some type of reference group is critical. Members of the reference group should be as much like those whose teachers participated in the program with respect to variables such as age, grade, sex, class, and school. While a one-to-one match is not necessary, the overall group profiles should be similar.

A comparison group is one in which students similar to students whose teachers participated in the program are identified. They are not necessarily randomly assigned as in control groups. Nevertheless, comparison students and experimental students should be administered the same instruments on the same schedule. One limitation of this study however, was that comparison students did not take the pretest. While the comparison group was formulated so as to match the characteristics of 700 randomly selected experimental students, the matching and hence, representativeness of the comparison group was less than ideal. For example, the proportions of seventh- and eighth-graders in each group seemed rather dissimilar; 27.6 percent of comparison students were seventh-graders as compared to 53.6 percent seventh-graders in the experimental group. The experimental group was relatively balanced with respect to grade, whereas the comparison group was comprised of proportionately more eighth-graders. Additionally, the comparison group students came from eight schools. Experimental group students were enrolled at only five of the eight comparison group schools. Therefore, there were students in the comparison group from three schools having no experimental students. Based on these descriptive data, equivalency of comparison group students and experimental group students at pretest time was tenuous.

One test that might have indicated some equivalency between groups at pretest time was the comparison of the seventh-grade comparison group posttest mean with the eighth-grade experimental group pretest mean. Presumably, end-of-year seventh graders would be similar in performance

to beginning-of-year eighth graders. Three two-tailed t-tests between means were performed. Tables 18 to 20 present the findings.

Table 18

t-Test Between Means of All Seventh-Grade Comparison Group Posttests And Eighth-Grade Experimental Pretest Only Subgroup (N=546)

Group	N	Mean	t	P
7th Grade Comparison Posttest Only Group	179	6.78	*3.18	.001
8th Grade Experimental Pretest Only Group	367	7.37		

*Approximation to t based on separate variances.

Table 19

t-Test Between Means of All Seventh-Grade Comparison Group Posttests And Eighth-Grade Experimental Pretest/Posttest Matched Pairs Pretests (N=1,002)

Group	N	Mean	t	P
7th Grade Comparison Posttest Only Group	179	6.78	*3.38	.001
8th Grade Experimental Matched Pairs Group	823	7.34		

*Approximation to t based on separate variances.

Table 20

t-Test Between Means of All Seventh-Grade Comparison Group Posttests and All Eighth-Grade Experimental Pretests (N=1,369)

Group	N	Mean	t	P
7th Grade Comparison Posttest Only Group	179	6.78	*3.57	.000
8th Grade Experimental All Pretest Groups	1,190	7.35		

*Approximation to t based on separate variances.

The null hypothesis that there is a difference between the seventh-grade comparison group posttest mean and the eighth-grade experimental group pretest mean was not rejected. Equivalency of the comparison and experimental at pretest time could not be established. It should be noted that experimental students took the pretest in December; not quite the beginning of the school year. Had the pretest been administered to experimental students in September, the t-tests performed might have indicated no difference between groups.

Keeping the design limitations in mind and not establishing equivalency of comparison and experimental groups at pretest time, comparisons between the comparison group and experimental group posttests were nevertheless, conducted. The null hypothesis that the experimental group posttest mean was less than or equal to the comparison group posttest mean was rejected at the .000 significance level for each grade individually and combined and for each experimental posttest subgroup. Such significant differences between comparison and experimental group posttest means would signify a true difference between the populations

and would indicate Writer's Clinic Program effectiveness. The uncertainty, however, as to whether or not the groups were equivalent at pretest time puts such interpretations into jeopardy.

The Relationship Between Holistic Essay Scores and Classroom Composition Grades and Teacher Ratings of Overall Writing Ability

The validity of an assessment instrument refers to the extent to which the instrument measures what it is intended to measure. This study attempted to determine the relationship between holistic essay scores and two contemporary criteria for writing: grades assigned to compositions written by students for English course requirements; and teacher ratings of overall student writing ability.

Teachers supplied composition grades for 215 students who had taken both the pretest and posttest for the Writer's Clinic evaluation. The grades were reported for the months of December through May. Additionally, teachers rated 173 of the pretest/posttest matched pairs students on overall writing ability both at the beginning of the year and at the end of the year on a scale of one to four (four was the highest rating).

Composition grades were split into two groups: those assigned in December, January and February were labeled as pretest essay grades; those assigned in March, April and May were labeled as posttest essay grades. Before determining whether or not there was a relationship between composition grades and essay test scores, t-tests between the pretest composition grades and posttest composition grades were run to determine whether or not student composition grades were higher towards

the end of the school term. Tables 21 through 23 show the t-test results for seventh-graders, eighth-graders and both grades combined, respectively.

Table 21

Correlated t-Test Between Means of Pretest and Posttest Composition Grades for Seventh-Grade Students (N=135)

Group	Mean	t	P
Pretest Composition Grades	2.90	4.58	.000
Posttest Composition Grades	3.07		

Table 22

Correlated t-Test Between Means of Pretest and Posttest Composition Grades for Eighth-Grade Students (N=80)

Group	Mean	t	P
Pretest Composition Grades	2.95	.62	.536
Posttest Composition Grades	2.98		

Table 23

Correlated t-Test Between Means of Pretest and Posttest Composition Grades Across Grade Levels (N=215)

Group	Mean	t	P
Pretest Composition Grades	2.92	4.09	.000
Posttest Composition Grades	3.04		

While there was not a significant difference between pre- and posttest composition grades for eighth-grade students, the null hypothesis of no difference across grades was rejected at the .000 level. Student essay grades were therefore, higher in general, at the end of the year than at the beginning of the year.

Correlated t-tests were also performed to determine whether or not a difference existed between teacher ratings of student writing ability at the beginning of the year and at the end of the year. These data are presented in tables 24 through 26.

Table 24

Correlated t-Test between Means of Pretest and Posttest Teacher Ratings of Overall Student Writing Ability for Seventh-Grade Students (N=116)

Group	Mean	t	P
December Ratings	2.51	7.13	.000
May Ratings	2.96		

Table 25

Correlated t-Test Between Means of Pretest and Posttest Teacher Ratings of Overall Student Writing Ability for Eighth-Grade Students (N=57)

Group	Mean	t	P
December Ratings	2.07	8.55	.000
May Ratings	2.84		

Table 26

Correlated t-Test Between Means of Pretest and Posttest Teacher Ratings of Overall Student Writing Ability Across Grade Levels (N=173)

Group	Mean	t	P
December Ratings	2.36	10.52	.000
May Ratings	2.92		

Clearly, teachers' ratings of student overall writing ability reflected better writing at the end of the year than in December. This indicates growth in student writing ability over time for seventh- and eighth-graders alike. The null hypothesis of no difference was rejected at the .000 level for pretest to posttest ratings within and across the grade levels.

As would be expected, pretest and posttest scores for the students for whom the validity criteria were supplied were significantly different; higher scores at posttest time. Tables 27 through 29 present the t-test results for the test variables for seventh-, eighth- and combined grade levels, respectively.

As was expected, posttest scores were significantly higher than pretest scores for both the seventh- and eighth-graders. The null hypothesis of no difference was rejected at the .007 level for seventh-graders and the .000 level for eighth-graders and both grades combined. In summary, these analyses showed teacher ratings, composition grades,

Table 27

Correlated t-Test Between Means of Seventh-Grade Pre- and Posttest Scores (N=135)

Group	Mean	t	P
Pretest	7.34	2.74	.007
Posttest	7.80		

Table 28

Correlated t-Test Between Means of Eighth-Grade Pre- and Posttest Scores (N=80)

Group	Mean	t	P
Pretest	6.96	4.43	.000
Posttest	8.05		

Table 29

Correlated t-Test Between Means of Pretest and Posttest Scores Across Grades (N=215)

Group	Mean	t	P
Pretest	7.20	4.93	.000
Posttest	7.89		

and essay test scores to reflect growth in writing over time.

Pearson correlation coefficients were calculated to determine the relationships between compositon grades and test scores and teacher ratings and test scores. Table 30 shows the means of each of the six validity component variables by grade. Correlation coefficients are presented in Table 31.

Table 30

Group Means by Validity Variables and by Grade

Variable	Scale		Grade		
	Low	High	7th	8th	7th & 8th
Pretest Compositions	1	5	2.90	2.95	2.92
Posttest Compositions	1	5	3.07	2.98	3.04
Teacher Ratings December	1	4	2.51	2.07	2.36
Teacher Ratings May	1	4	2.96	2.84	2.91
Pretest Essay Scores	3	12	7.34	6.96	7.20
Posttest Essay Scores	3	12	7.80	8.05	7.89

The Pearson correlation coefficients indicate there are relationships between holistic essay test scores and teacher ratings of overall writing ability and composition grades. The strongest relationship lies in the correlation between teacher ratings and holistic essay scores; .26 for December ratings and .53 for May ratings. The lower correlation with December ratings rests with the fact that teachers rated the students

Table 31

Pearson Correlation Coefficients

Criterion Variables	7th Grade		8th Grade		Across Grades	
	Pretest	Posttest	Pretest	Posttest	Pretest	Posttest
Teacher Ratings December	.25***	--	.15	--	.26***	--
Teacher Ratings May	--	.55***	--	.49***	--	.53***
Pretest Composition Grades	.03	--	.43***	--	.16**	
Posttest Composition Grades	--	.02	--	.34***	--	.13*

Significance Levels: * $p < .05$
** $p < .01$
*** $p < .001$

in May (they were asked to recall student writing ability as it was five months earlier and as it was in May). It seems reasonable to assume that composition grades were referred to in order to rate the students' ability as it was at pretest time. The lower correlation between holistic essay test scores and composition grades is therefore, reflected in the December ratings. Lower correlations of test scores with grades also seems likely due to the fact that grades are distinctly different from scores; they take into account individual abilities, the topic responded to and the structure of the classroom curriculum. Still, the criteria of both composition grades and teacher ratings were related to obtained holistic essay scores and therefore, serve as acceptable predictors of concurrent validity to some extent.

Summary of Findings

The distribution of test scores for each of the three readings and summed totals approximated the normal distribution; the theoretical model of holistic essay scoring was supported by the case study. While tests of significance were not performed, the null hypothesis that the distribution of scores did not approximate the normal distribution was rejected with some certainty.

Discrepancies in scoring among the three readers per essay paper were found in 4.46 percent of the 5,788 papers included for analyses. Relatively strong agreement between readers was further substantiated by reader relaibility coefficients ranging from .77 to .84. This reading reliability increased confidence in the validity of the assigned scores and enabled further analyses of the data.

Within each of the pretest and posttest essays, eighth-grade students consistently performed better in terms of total essay scores. The null hypothesis of no difference in scores between the grade levels was rejected beyond the .05 level of significance.

Growth in writing ability over a five-month period was substantiated. The null hypothesis of no difference between pretest and posttest mean scores within the experimental group was rejected at the .000 significance level for seventh- and eighth-grades individually and combined.

Determining the effectiveness of the teacher inservice program on writing was difficult to establish primarily because comparison (no treatment) students did not take the pretest and it could not be said with any degree of certainty that this group was similar to the experimental (treatment) group at pretest time. Alternate analyses conducted

to try to establish equivalency of comparison and experimental group students at pretest time were not successful. Keeping the design limitations in mind, posttest comparisons made between the treatment and no treatment groups indicated a difference significant beyond the .001 level; experimental students' posttest scores being higher than comparison group students' scores.

The study also attempted to determine the relationship of holistic essay scores with student class composition grades and teacher ratings of student writing ability. Relationships were established by Pearson correlation coefficients. The strongest relationship was between posttest holistic essay scores and teacher ratings of student writing ability in May (r=.53 across grades). The weakest relationship was between holistic essay scores and posttest period composition grades (r=.13 across grades).

CONCLUSIONS

Within the limitations of the study, several conclusions are warranted. First, the holistic method used at the school district level for evaluating student writing and a program designed to heighten that ability, seemed to be an efficient means of essay scoring. In a two-day period 5,788 student essay examinations were each scored three times by 44 readers. While not as fast as machine scoreable multiple-choice test (indirect measurement) answer sheets, the readers essentially read 17,364 papers in approximately six hours; an average of about 64 papers scored per hour per reader.

The ease of the scoring in terms of time along with the relative reliability of the scoring, contributed to the efficiency and

hence, utility of employing the holistic technique for the evaluation of writing. Holistic scoring as it was used in the study, provided a means of measurement that was able to discriminate among the good writers and the poorer writers. Hypothesized results, including a difference in performance of seventh-graders and eighth-graders and growth in writing over time were substantiated by the significant differences between grade levels between and within pretest and posttest essay administrations.

Content and construct validity of the essay examination had been established prior to the essay scoring by clinicians of the Writer's Clinic program. The study was able to support some current validity by finding relationships between the obtained essay scores and teacher ratings of writing ability. Procedures implemented for the scoring session confirmed the notion of ranking as did the teacher ratings; ranking in the sense of judging writing for what it is rather than for what it should be. The scoring therefore, followed the theoretical model of holistic scoring; judging a paper against others rather than against a preconceived ideal.

Teacher acceptance of the holistic method, while not directly analyzed for this study, is of vital importance to the utility of the method. Utility of the method for the stated purpose of directly evaluating student composition was supported by the interest of the teacher group. The involved teachers, at first perhaps skeptical of the method, accepted it as a useful tool with additional applications for the classroom.

EDUCATIONAL IMPLICATIONS

The findings of this study indicate the holistic method of scoring provides one solution to the problem of evaluating writing directly, efficiently, and reliably. Holistic evaluation provides a starting point for evaluation. While readers may not be able to use the method for individual diagnoses, they are able to identify general problems in their students' writing. They see the writing as it is and can begin to view writing in its totality; to view writing as a whole rather than for individual aspects of writing skill.

Faced with impressions of weaknesses or strengths of student writing in general, the reader can gear instructional objectives to improve students' abilities in weak areas as well as to build on the areas of strength. Looking at writing as it is, and working with students were difficulties lie, can bring the students closer to the preconceived ideal, closer to what writing truly should be.

Applications of the method for the classroom have few boundaries. The ease and quickness of the holistic method provides the teacher with a tool so that more writing assignments can be given in order to provide the student practice in writing and to provide the student the opportunity to express himself or herself by the written word.

Essay examinations will doubtfully ever replace multiple-choice tests of writing skill. Direct measures however, will continue to supplement indirect measures of writing ability. Likewise, holistic scoring can never replace, but can provide a supplement to, analytic scoring. And the total approach to analyzing writing can be of value not only for national standardized essay examinations but also for use

within the individual classroom, within and across subject area departments, within the school, and within the school district.

It is a fundamental requirement in a modern society for its citizenry to possess a range of communication skills; writing being one of the most basic. Thus, continued research in assessing student writing skills is necessary to assist administrators and teachers in elementary and secondary education. The goal of such research is quite direct - to encourage thought and its expression.

Reference:

Godshalk, F. I., Swineford, F., & Coffman, W. E. The Measurement of Writing Ability. New York: CEEB, 1966.